

First we will consider the case where  $\mu_a$  is estimated with a finite-dimensional parametric model, i.e., it is assumed that

$$\mu_a(x) = \mu_a(x; \beta)$$

for some real-valued parameter  $\beta \in \mathbb{R}^p$ . A prominent example might include a logistic regression model  $\mu_a(x; \beta) = \text{expit}(\beta_0 + \beta_1 a + \beta_2^T x)$ . The parameter  $\beta$  could be estimated via maximum likelihood or some relevant m-estimator, for example. Note that the discrete covariate setup can be viewed as a special case of this, since then for each  $a = 0, 1$  the function  $\mu_a(x)$  can be estimated with a saturated model with  $d$  parameters (e.g.,  $d - 1$  level indicators and an intercept).

In fact, we have already analyzed the estimator (4.2) in the case where it is assumed that  $\mu_a$  follows a parametric model, in Theorem 3.1. Recall when we analyzed the parametric plug-in estimator we did not rely on the randomization at all, so the same analysis applies here. The relevant theorem is repeated below for posterity.

**Theorem 4.2.** *Let  $f(x) = \mu_1(x) - \mu_0(x)$  and  $\mu_a(x) = \mathbb{E}(Y \mid X = x, A = a)$ , so that  $\psi = \mathbb{E}\{f(X)\}$  is the average treatment effect. Assume the parametric model*

$$\mu_a(x) = \mu_a(x; \beta_a)$$

*for some  $\beta = (\beta_0, \beta_1) \in \mathbb{R}^p$ . Suppose the estimator  $\hat{\beta} \in \mathbb{R}^p$  solves an estimating equation so that*

$$\mathbb{P}_n\{m(Z; \hat{\beta})\} = 0.$$

*Assume  $m(z; \beta) \in \mathbb{R}^p$  is Lipschitz in  $\beta$ , and that  $\mathbb{E}\{m(z; \beta)\}$  is differentiable at the true  $\beta$  satisfying  $\mathbb{E}\{m(Z; \beta)\} = 0$  with nonsingular derivative matrix. Then*

$$\hat{\psi}_{reg} - \psi = (\mathbb{P}_n - \mathbb{P})g(Z; \beta) + o_{\mathbb{P}}(1/\sqrt{n})$$

*where*

$$g(z; \beta) = f(x; \beta) + \frac{\partial \mathbb{E}\{f(X; \beta)\}}{\partial \beta^T} \left\{ \frac{\partial \mathbb{E}\{m(Z; \beta)\}}{\partial \beta^T} \right\}^{-1} m(z; \beta)$$

*and so is root- $n$  consistent and asymptotically normal.*

Theorem 3.1 shows that if a correct parametric model is available for the regression functions  $\mu_a$ , and under some regularity conditions (essentially smoothness of the model), then the resulting plug-in regression estimator is root- $n$  consistent and asymptotically normal. Confidence intervals can be constructed using an estimate of the closed-form asymptotic variance  $\widehat{\text{var}}\{\hat{g}(Z; \hat{\beta})\}$ , or via the bootstrap which is typically easier. Bootstrap estimates would be computed in the usual way: take a bootstrap sample (i.e., sample  $n$  observations with replacement), re-estimate the regression functions  $\mu_a$ , and construct the plug-in estimator (and then repeat many times).

In practice, there is often not enough detailed background knowledge to justify a particular parametric model, especially in the presence of continuous covariates. This motivates the use of a nonparametric version of the regression estimator (4.2). However, in the nonparametric case, the analysis from Theorem 3.1 no longer applies, since  $\mu_a$  is no longer assumed to be indexed by a finite-dimensional parameter. This means that the term  $T_2 = \mathbb{P}(\hat{f} - f)$  in the proof cannot be analyzed by simply differentiating with respect to  $\beta$ .

In fact, our discussion from Section 3.3.3 concerning nonparametric covariate adjustment in experiments applies as well, since as in the parametric case we did not rely on randomization there. To reiterate: when the regression functions  $\mu_a$  are estimated nonparametrically, the term  $T_2 = \mathbb{P}(\hat{f} - f)$  cannot in general be expected to be  $O_{\mathbb{P}}(1/\sqrt{n})$ , so that the regression estimator would typically inherit slow convergence rates from estimating  $\mu_a$  flexibly. This means that a nonparametric version of the regression estimator (4.2) has advantages over a parametric version because it will generally be consistent under weaker conditions, and thus more robust; however it has a potential disadvantage in that it will be less efficient, compared to if the parametric model is correctly specified. Further, it is not only slow rates that complicate the use of this estimator: for general nonparametric estimators it will often not have a tractable limiting distribution, and even when it does it would typically not be correctly centered (without undersmoothing), and so inference would be complicated at best.

Importantly, however, there are some exceptions to this story. Namely, if one specifies particular nonparametric estimators  $\hat{\mu}_a$  (e.g., based on kernels, splines, or nearest-neighbor regression), undersmooths (so that smaller bias of  $\hat{\mu}_a$  is traded off for larger variance, rather than the usual balancing), and makes some particular structural assumptions (e.g., that the regression functions are Hölder smooth), then nonparametric regression estimators can be root-n consistent (and sometimes asymptotically normal).

We will first discuss such an analysis of matching estimators, and then briefly mention series/spline estimators.

## Matching

Matching is often treated as an entirely different adjustment approach, but it can be viewed as a particular nonparametric regression-based estimator, for example using k-nearest neighbor regression to estimate  $\mu_a$ . Following Abadie and Imbens [2006], one simple version of matching uses  $\hat{\psi}_{reg}$  from (4.2) with

$$\hat{\mu}_{A_i}(x_i) = Y_i \quad , \quad \hat{\mu}_{1-A_i}(x_i) = \frac{1}{K} \sum_{j \in \mathcal{J}_K(i)} Y_j$$

where  $\mathcal{J}_K(i)$  is the set of  $K$  indices with the treatment opposite of  $a_i$ , who are closest in terms of covariates, i.e.,  $\mathcal{J}_K(i) = \{j_1(i), \dots, j_K(i)\}$  for

$$j_k(i) = \left\{ j : \sum_{\ell: A_\ell \neq A_i} \mathbb{1}\{\|X_\ell - X_i\|_2 \leq \|X_j - X_i\|_2\} = k \right\}$$

the index of the  $k^{th}$  closest match, for  $\|\cdot\|_2$  the Euclidean norm. This corresponds to matching with replacement, since the same indices can reappear in  $\mathcal{J}_K(i)$  for different subjects  $i$ . Of course, other schemes can also be used, e.g., based on matching without replacement, or via other distance metrics, or using a variable number of neighbors  $k$ , etc.

Abadie and Imbens [2006] show that, under usual iid sampling setup and consistency/exchangeability/positivity assumptions, and if

- the covariates  $X \in \mathbb{R}^d$  are continuous, have compact and convex support, and density bounded away from zero, and if
- $\mu_a(x)$  is Lipschitz in  $x$  for  $a \in \{0, 1\}$ ,

then for fixed  $K$  the matching estimator has (conditional) bias

$$\mathbb{P}_n \left( (2A - 1) \left[ \mu_{1-A}(X) - \mathbb{E}\{\hat{\mu}_{1-A}(X) \mid X, A\} \right] \right) = O_{\mathbb{P}}(1/n^{1/d})$$

whose expected value is not of smaller order than  $O(1/n^{2/d})$ . This means that the estimator  $\hat{\psi}_{reg}$  when based on  $K$ -nearest neighbor matching is

- root-n consistent and asymptotically normal, if  $d = 1$ ;
- root-n consistent but not necessarily asymptotically normal, if  $d = 2$ ;
- not root-n consistent or asymptotically normal, if  $d > 2$ .

This illustrates the kind of result we discussed above heuristically: in general nonparametric versions of the regression estimator  $\hat{\psi}_{reg}$  are not expected to be root-n consistent and asymptotically normal, but some exceptions can occur, e.g., for particular  $K$ -nearest neighbor estimators, under smoothness conditions and/or low dimensions.

There are other examples of nonparametric estimators being root-n consistent with undersmoothing. Hahn [1998] has a nice paper that explores efficiency bounds for average treatment effects, and considers undersmoothed regression estimators based on series/spline regression. He shows that: if one uses particular orthonormal bases, and carefully tuned number of series terms, and if the propensity score and regression functions are infinitely differentiable, then a regression-type estimator is root-n consistent and asymptotically normal. Some caveats of these kinds of analyses are that in practice:

- it is difficult to construct bases satisfying the required regularity conditions;
- it can be even more difficult to pick the right number of basis terms (i.e., it requires undersmoothing – suboptimal tuning by trading less bias for more variance).

### 4.4.3 Weighting

The inverse-probability-weighted estimator (4.3) can be motivated from importance sampling or representativeness arguments, and depends on an estimate of the unknown propensity score  $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$ . In fact, its analysis is essentially the same as that of the regression-based estimator; the main difference is we define

$$\hat{f} = \left\{ \frac{A}{\hat{\pi}(X)} - \frac{1-A}{1-\hat{\pi}(X)} \right\} Y$$

instead of  $\hat{f} = \hat{\mu}_1 - \hat{\mu}_0$ .

First we will consider the case where the propensity score  $\pi$  is estimated with a parametric model, i.e.,

$$\hat{\pi}(x) = \pi(x; \hat{\beta}).$$

For example with logistic regression we might have  $\pi(x; \beta) = \text{expit}(\beta_0 + \beta_1^T x)$ . As with the regression estimator  $\hat{\psi}_{reg}$ , the parameter  $\beta$  could be estimated via maximum likelihood or m-estimation. In the next result we give an analog of Theorem 3.1 for the parametric weighting estimator. Note that this setup can be viewed as a semiparametric model in which the propensity score  $\pi$  is restricted to follow a known parametric form, but the regression functions  $\mu_a$  are left unrestricted.

**Theorem 4.3.** *Let  $f(z) = \left\{ \frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)} \right\} y$  and  $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$ , so that  $\psi = \mathbb{E}\{f(Z)\}$  is the average treatment effect. Assume the parametric model*

$$\pi(x) = \pi(x; \beta)$$

*for some  $\beta \in \mathbb{R}^p$ . Suppose the estimator  $\hat{\beta} \in \mathbb{R}^p$  solves an estimating equation so that*

$$\mathbb{P}_n\{m(Z; \hat{\beta})\} = 0.$$

*Assume  $m(z; \beta) \in \mathbb{R}^p$  is Lipschitz in  $\beta$ , and that  $\mathbb{E}\{m(z; \beta)\}$  is differentiable at the true  $\beta$  satisfying  $\mathbb{E}\{m(Z; \beta)\} = 0$  with nonsingular derivative matrix. Then*

$$\hat{\psi}_{ipw} - \psi = (\mathbb{P}_n - \mathbb{P})g(Z; \beta) + o_{\mathbb{P}}(1/\sqrt{n})$$

*where*

$$g(z; \beta) = f(z; \beta) + \frac{\partial \mathbb{E}\{f(Z; \beta)\}}{\partial \beta^T} \left\{ \frac{\partial \mathbb{E}\{m(Z; \beta)\}}{\partial \beta^T} \right\}^{-1} m(z; \beta)$$

*and so is root-n consistent and asymptotically normal.*

*Proof.* By Lemma 3.1 we have

$$\hat{\psi} - \psi = Z^* + T_1 + T_2$$

where  $Z^* = (\mathbb{P}_n - \mathbb{P})f$  and  $T_1$  and  $T_2$  defined accordingly. By Lemma 3.2 we have

$$\hat{\beta} - \beta = (\mathbb{P}_n - \mathbb{P}) \left[ \left\{ \frac{\partial \mathbb{E}(m(Z; \beta))}{\partial \beta^T} \right\}^{-1} m(Z; \beta) \right] + o_{\mathbb{P}}(1/\sqrt{n})$$

which also is enough to imply  $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$ . Further by the delta method we have

$$\begin{aligned} T_2 &= \mathbb{P}(\hat{f} - f) = h(\hat{\beta}) - h(\beta) \\ &= (\mathbb{P}_n - \mathbb{P}) \left[ \frac{\partial h(\beta)}{\partial \beta^T} \left\{ \frac{\partial \mathbb{E}(m(Z; \beta))}{\partial \beta^T} \right\}^{-1} m(Z; \beta) \right] + o_{\mathbb{P}}(1/\sqrt{n}) \end{aligned}$$

for  $h(\beta) = \mathbb{E}\{f(Z; \beta)\} = \mathbb{E}\left[\left\{\frac{A}{\pi(X; \beta)} - \frac{1-A}{1-\pi(X; \beta)}\right\}Y\right]$ . Combining terms gives the result.  $\square$

The analysis and interpretation of the weighting estimator parallels that of the regression estimator. In particular, Theorem 4.3 shows that if a correct parametric model is available for the propensity scores  $\pi$ , and under some regularity conditions (essentially smoothness of the model), then the resulting inverse-probability-weighted estimator is root-n consistent and asymptotically normal. Confidence intervals can be constructed using an estimate of the closed-form asymptotic variance  $\widehat{\text{var}}\{\hat{g}(Z; \hat{\beta})\}$ , or via the bootstrap. Surprisingly, another variance estimator is available for the weighting estimator: namely, valid but conservative inference can be obtained via the variance estimate  $\widehat{\text{var}}\{f(Z; \hat{\beta})\}$ , which is the variance estimate that would be used if the propensity scores were known to be equal to  $\hat{\pi}(x) = \pi(x; \hat{\beta})$  [Tsiatis, 2006]

As with regression estimators, the analysis for weighting estimators also needs to be amended in the nonparametric case. Here, the term  $T_2$  can essentially inherit the convergence rate of  $\hat{\pi}$  in the sense that

$$\begin{aligned} \mathbb{P}(\hat{f} - f) &= \mathbb{P} \left\{ \left( \frac{\pi - \hat{\pi}}{\hat{\pi}} \right) \mu_1 - \left( \frac{\hat{\pi} - \pi}{1 - \hat{\pi}} \right) \mu_0 \right\} \\ &= \mathbb{P} \left\{ (\pi - \hat{\pi}) \left( \frac{\mu_1}{\hat{\pi}} + \frac{\mu_0}{1 - \hat{\pi}} \right) \right\} \\ &\lesssim \mathbb{P}|\hat{\pi} - \pi| \leq \sqrt{\mathbb{P}\{(\hat{\pi} - \pi)^2\}} = \|\hat{\pi} - \pi\| \end{aligned}$$

where the second inequality follows by Cauchy-Schwarz. Thus we should again not expect root-n rates for nonparametric weighting estimators, and confidence intervals (even using bootstrap) may not be correctly centered, and thus invalid.

However, again similar to the regression case, Hirano et al. [2003] show that undersmoothing can correct these issues, under some assumptions. In particular, they show that under some relatively strong smoothness conditions on the propensity score  $\pi$  (i.e., that  $\pi$  has seven times as many derivatives as it does dimensions  $d$ ), and undersmoothing, a series-based nonparametric inverse-probability-weighted estimator is root-n consistent and asymptotically normal.

#### 4.4.4 Doubly Robust Estimation

Although the doubly robust estimator (4.4) is perhaps less intuitive than the regression and inverse-probability-weighted estimators, it comes with important advantages. The classic advantage is that, when the nuisance functions  $\pi$  and  $\mu_a$  are estimated with parametric models, the doubly robust estimator is root-n consistent and asymptotically normal even if one of the two models is completely misspecified. This classic result is given in the following theorem.

**Theorem 4.4.** *Let  $f(z) = \left\{ \frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)} \right\} \{y - \mu_a(x)\} + \mu_1(x) - \mu_0(x)$ , so that  $\psi = \mathbb{E}\{f(Z)\}$  is the average treatment effect. Assume at least one of the parametric models*

$$\pi(x) = \pi(x; \alpha), \quad \mu_a(x) = \mu_a(x; \beta_a)$$

*for some  $\eta = (\alpha, \beta_0, \beta_1) \in \mathbb{R}^p$ . Suppose the estimators  $\hat{\eta} \in \mathbb{R}^p$  solve an estimating equation so that*

$$\mathbb{P}_n\{m(Z; \hat{\eta})\} = 0.$$

*Assume  $m(z; \eta) \in \mathbb{R}^p$  is Lipschitz in  $\eta$ , and that  $\mathbb{E}\{m(z; \eta)\}$  is differentiable at the true  $\eta$  satisfying  $\mathbb{E}\{m(Z; \eta)\} = 0$  with nonsingular derivative matrix. Then*

$$\hat{\psi}_{dr} - \psi = (\mathbb{P}_n - \mathbb{P})g(Z; \eta) + o_{\mathbb{P}}(1/\sqrt{n})$$

where

$$g(z; \eta) = f(z; \eta) + \frac{\partial \mathbb{E}\{f(Z; \eta)\}}{\partial \eta^T} \left\{ \frac{\partial \mathbb{E}(m(Z; \eta))}{\partial \eta^T} \right\}^{-1} m(z; \eta)$$

and so is root-n consistent and asymptotically normal.

*Proof.* The proof is the same as Theorem 4.3, replacing  $f$  accordingly.  $\square$

**Remark 4.5.** When both the propensity score and regression models are correctly specified, it follows that

$$\hat{\psi}_{dr} - \psi = (\mathbb{P}_n - \mathbb{P})f(Z) + o_{\mathbb{P}}(1/\sqrt{n})$$

since in that case  $\frac{\partial \mathbb{E}\{f(Z; \eta)\}}{\partial \eta^T} = 0$ .

Theorem 4.4 shows the surprising phenomenon that the doubly robust estimator remains root-n consistent even if one of the two nuisance estimators is misspecified, giving the analyst “two chances” at fast rates and valid inference.

However, a skeptic might argue that the chance of correctly specifying a parametric model is essentially zero, in which case the chance of correctly specifying one of two parametric models is twice zero, and still zero. In fact the doubly robust estimator has an even more useful advantage. Namely, it can be root-n consistent and asymptotically normal even when the nuisance functions  $\pi$  and  $\mu_a$  are estimated flexibly at slower than root-n rates, in a wide variety of settings. Surprisingly, this can be true even without

committing a priori to particular estimators or function classes, and without using any undersmoothing tricks. Contrast this with the estimators (4.2) and (4.3), which would generally not be root-n consistent when the nuisance functions are estimated at nonparametric rates, without requiring careful analysis of particular undersmoothed estimators.

**Theorem 4.5.** *Let  $f(z) = \left\{ \frac{a}{\pi(x)} - \frac{1-a}{1-\pi(x)} \right\} \{y - \mu_a(x)\} + \mu_1(x) - \mu_0(x)$ , so that  $\psi = \mathbb{E}\{f(Z)\}$  is the average treatment effect. Assume  $\|\hat{f} - f\| = o_{\mathbb{P}}(1)$  and either:*

1.  *$f$  and its estimate  $\hat{f}$  are contained in a Donsker class, or*
2. *the estimate  $\hat{f}$  is constructed from a separate independent sample.*

*Also assume  $\mathbb{P}(\hat{\pi} \in [\epsilon, 1 - \epsilon]) = 1$ . Then if  $\|\hat{\pi} - \pi\| \sum_a \|\hat{\mu}_a - \mu_a\| = o_{\mathbb{P}}(1/\sqrt{n})$  it follows that*

$$\hat{\psi}_{dr} - \psi = (\mathbb{P}_n - \mathbb{P})f(Z) + o_{\mathbb{P}}(1/\sqrt{n})$$

*and so is root-n consistent and asymptotically normal.*

*Proof.* By Lemma 3.1 we have

$$\hat{\psi} - \psi = Z^* + T_1 + T_2$$

where  $Z^* = (\mathbb{P}_n - \mathbb{P})f$  and  $T_1$  and  $T_2$  are defined accordingly. The consistency of  $\hat{f}$  in  $L_2$  norm, together with either the Donsker condition or sample splitting, ensures that  $T_1 = o_{\mathbb{P}}(1/\sqrt{n})$ . Finally for  $T_2 = \mathbb{P}(\hat{f} - f)$  and  $f = f_1 - f_0$  we have

$$\begin{aligned} \mathbb{P}(\hat{f}_1 - f_1) &= \mathbb{P} \left\{ \frac{A}{\hat{\pi}} (Y - \hat{\mu}_A) + \hat{\mu}_1 - \mu_1 \right\} \\ &= \mathbb{P} \left\{ \left( \frac{\pi}{\hat{\pi}} - 1 \right) (\mu_1 - \hat{\mu}_1) \right\} \\ &\leq \frac{1}{\epsilon} \mathbb{P} \left\{ \left| \pi - \hat{\pi} \right| |\mu_1 - \hat{\mu}_1| \right\} \\ &\leq \frac{1}{\epsilon} \|\hat{\pi} - \pi\| \|\hat{\mu}_1 - \mu_1\| \end{aligned}$$

where the second line follows by iterated expectation, the third by the bound on  $\hat{\pi}$ , and the fourth by Cauchy-Schwarz. The exact same logic shows that  $\mathbb{P}(\hat{f}_0 - f_0) \leq \frac{1}{1-\epsilon} \|\hat{\pi} - \pi\| \|\hat{\mu}_0 - \mu_0\|$ . Therefore  $T_2 = o_{\mathbb{P}}(1/\sqrt{n})$  and the result follows.  $\square$

*Remark 4.6.* Note that Theorem 4.5 tells us that the doubly robust estimator is root-n consistent and asymptotically normal for a wide range of nuisance estimators  $(\hat{\pi}, \hat{\mu}_a)$ . For example, a sufficient condition for the product of  $L_2$  norms being  $o_{\mathbb{P}}(1/\sqrt{n})$  is that both  $\hat{\pi}$  and  $\hat{\mu}_a$  converge at (faster than)  $n^{-1/4}$  rates. This could be satisfied for Hölder functions in  $\mathcal{H}(s)$  if  $s > d/2$ , or for  $s$ -sparse functions if  $s \log(d/s) < \sqrt{n}$ . However, any product of the order  $o_{\mathbb{P}}(1/\sqrt{n})$  would result in root-n consistency and asymptotic normality: for example if one nuisance function was estimated at root-n rates, the other would only need to be estimated consistently, at any rate.

# Appendix A

## Notation Guide

$Y^a$	Potential outcome under treatment/exposure $A = a$
$\perp\!\!\!\perp$	Statistically independent
$\xrightarrow{p}$	Convergence in probability
$\rightsquigarrow$	Convergence in distribution
$O_{\mathbb{P}}(1)$	Bounded in probability
$o_{\mathbb{P}}(1)$	Converging in probability to zero
$\mathbb{P}_n$	Sample average operator, as in $\mathbb{P}_n(\hat{f}) = \mathbb{P}_n\{\hat{f}(Z)\} = \frac{1}{n} \sum_{i=1}^n \hat{f}(Z_i)$
$\mathbb{P}$	Conditional expectation given the sample operator, as in $\mathbb{P}(\hat{f}) = \int \hat{f}(z) d\mathbb{P}(z)$
$\ \cdot\ $	$L_2(\mathbb{P})$ norm $\ f\  = \sqrt{\mathbb{P}(f^2)}$ or Euclidean norm, depending on context
$\ \cdot\ _1$	$L_1(\mathbb{P})$ norm $\ f\ _1 = \mathbb{P}( f )$
$\ \cdot\ _{\infty}$	$L_{\infty}$ or supremum norm $\ f\ _{\infty} = \sup_z  f(z) $
$\mathcal{H}(s)$	Hölder class of functions with smoothness index $s$
$\lesssim$	Less than or equal, up to a constant multiplier



# Bibliography

- A. Abadie and G. W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.
- P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.
- H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.
- P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press, 1993.
- D. D. Boos and L. A. Stefanski. *Essential Statistical Inference: Theory and Methods*. New York: Springer, 2013.
- V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. M. Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.
- W. G. Cochran. *Sampling Techniques*. John Wiley & Sons, 1977.
- M. Davidian, A. A. Tsiatis, and S. Leon. Semiparametric estimation of treatment effect in a pretest–posttest study with missing data. *Statistical Science*, 20(3):261, 2005.
- L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.
- D. A. Freedman. Randomization does not justify logistic regression. *Statistical Science*, pages 237–249, 2008.
- S. Greenland, J. M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science*, pages 29–46, 1999.
- L. Györfi, M. Kohler, A. Krzykacz, and H. Walk. *A Distribution-Free Theory of Non-parametric Regression*. Springer, 2002.
- J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

## BIBLIOGRAPHY

---

- K. Hirano, G. W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.
- P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.
- G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.
- G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
- E. H. Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.
- E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Nonparametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B*, 79(4):1229–1245, 2017.
- E. H. Kennedy, S. Balakrishnan, and M. G’Sell. Sharp instruments for classifying compliers and generalizing causal effects. *The Annals of Statistics (to appear)*, 2019a.
- E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019b.
- S. Leon, A. A. Tsiatis, and M. Davidian. Semiparametric estimation of treatment effect in a pretest-posttest study. *Biometrics*, 59(4):1046–1055, 2003.
- D. Michaels. *Doubt is their product: how industry’s assault on science threatens your health*. Oxford University Press, 2008.
- J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.
- J. Pfanzagl. *Contributions to a general asymptotic statistical theory*, volume 13. Springer, 1982.
- J. M. Robins and A. Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.
- J. M. Robins and A. Rotnitzky. Comments on: Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:920–936, 2001.
- J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87(1):113–124, 2000.

## BIBLIOGRAPHY

---

- J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.
- J. M. Robins, A. Rotnitzky, and L. P. Zhao. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121, 1995.
- D. B. Rubin and M. J. van der Laan. Empirical efficiency maximization: Improved locally efficient covariate adjustment in randomized experiments and survival analysis. *The International Journal of Biostatistics*, 4(1), 2008.
- Z. Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010.
- A. A. Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.
- A. B. Tsybakov. *Introduction to Nonparametric Estimation*. New York: Springer, 2009.
- M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer, 2003.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.
- L. Yang and A. A. Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.
- M. Zhang, A. A. Tsiatis, and M. Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3):707–715, 2008.